World Journal of Information Technology
ISSN: 2959-9911
DOI: https://doi.org/10.61784/wjit3039

THE PREDICTION OF OLYMPIC MEDAL TABLE BASED ON LINEAR REGRESSION MODELING

Lei Zhao
School of Mechanical and Electrical Engineering, Anhui University of Science and Technology, Huainan 232001, Anhui, China.
Corresponding Email: Zl897932@163.com

Abstract: As the world's most influential sporting event bringing together the world's best athletes, the Olympic Games is the highest stage for competitive sports. It inspires more people to participate. In this paper, in order to predict the total medals and gold medals won by each country in the 2028 Olympic Games, a multiple linear regression model is constructed by considering the historical medal datas, the number of athletes' participation and the types and number of participating events and other characteristic variables as the indexes, and takes the evaluation coefficient $R^2$and the mean squared error MSE as the model evaluation indexes. Through the established medal list and historical trend, the prediction interval of total medals and the prediction interval of gold medals are analyzed, and those countries that may progress or regress in the 2028 Olympic Games are analyzed and obtained, and by the prediction this paper selects the top 10 progressing countries and the 10 countries with obvious regression. In addition, the prediction of those countries that have never won a prize is also made to explore the possibility of winning a medal, and for this purpose, this paper adopts a binary classification model and logistical regression model, and the probabilities of winning a first medal are obtained by selecting the data of the countries that have never won the award.
Keywords: Predicting the number of medals; Linear regression model; Model evaluation; Binary classification model and logistical regression model

1 INTRODUCTION

As the largest and most influential comprehensive sports event in the world, the Olympic Games has an important significance to the global society, economy, culture and other aspects that can not be ignored. The prediction of the medal table has always been the focus of public and professional attention[1]. In existing studies, time series models , such as gray theory prediction model and stochastic time series analysis model[2], are mainly used, which rely heavily on the quality of finite historical data, and missing or abnormalities will affect the prediction, and can not deal with nonlinear relationships.

And according to the national economic level and the total population to establish the econometric prediction model to predict the number of medals [3], can reflect the degree of influence of each factor size. But it has limitations tend to ignore the traditional strengths of each country's sports programs, usually lacks of knowledge of the specifics of a competition, athletes' past performance and psychological quality.

Neural network nonlinearity was used to fit and predict the number of medals by quantitatively predicting and studying the GDP per capita of each country as well as the previous medal scores[4]. neural network is able to efficiently learn complex nonlinear relationships in the data, and through the multilayer structure (especially deep neural networks), it can abstract features of the data can be learned, which is suitable for dealing with tasks with complex patterns , but is prone to overfitting in the simulation process, especially if the training data is insufficient or the model is too complex.

By inputting several independent variables such as the history of awards for each of its countries, the number of participating athletes, and outputting the dependent variable, the number of medals, Support Vector Machines (SVMs) are able to handle nonlinearly differentiable problems by introducing kernel functions to map the original features to a higher dimensional space. This enables SVMs to handle complex nonlinear classification problems, however, the performance of SVMs depends heavily on the choice of parameters, such as the regularization parameter C and the kernel function. Different combinations of parameters and kernel functions can significantly affect the performance of the model and need to be tuned by methods such as cross-validation. Therefore it should be simplified according to the measurement method. And it is sensitive to large-scale data and feature items.

So unlike those that use time series model, this paper constructs a comprehensive multiple linear regression model[5], which is suitable to be used with sufficiently large amount of data and suitable choice of independent variables to provide its reliable prediction results[6]. Factors affecting the number of medals are entered as characteristic variables to predict the performance of countries in the 2028 Olympic Games, especially the number of gold medals. It provides its reliable predictions with enough data and suitable choice of independent variables. It also provides an in-depth analysis of the various factors affecting these predictions. This model is able to handle the relationship between multiple independent variables and a dependent variable, captures the joint influence of multiple independent variables on the dependent variable, and fits the data better by analyzing the regression coefficients, which allows for the determination of the degree of influence of each of the independent variables on the corresponding variable, and is insensitive to small variations in the data.

2 FORECASTING THE OLYMPIC MEDAL TABLE RESEARCH METHOD AND MODEL CONSTRUCTION

In this paper through a multi-dimensional analysis, first collect the raw data that need to be analyzed and observe the laws behind the phenomenon from the data to obtain valid data, then select independent variables such as historical medal data of each country, the number of athletes participating and the type of events participating etc, what's more, test whether there is a linear relationship between the independent variables and the dependent variable, finally determine the dependent variable Y (the predicted total number of medals or gold medals) with the independent variables X_1, X_2,...,X_n of the linear relationship to establish a linear multiple regression model and comprehensively explore the prediction of the number of medals in the 2028 Los Angeles Olympic Games and the analysis of influencing factors. On this basis, the quantity R^2 and the mean square error MSE are calculated. In this paper, the countries that may progress and regress in the 2028 Olympic Games are analyzed based on the previous historical total medals data and gold medals data(based on the data provided by The 2025 Mathematical Contest in Modeling and the download website is http://www.mcmcontest.com) and the predicted medal table. The article is tightly structured, and each part of the paper starts from a different perspective to provide rich theoretical support and empirical analysis for the prediction models.

2.1 Modeling Establishment

2.1.1 Data normalization

After pre-processing the different kinds of data, in order to eliminate inconsistencies between the different variables. Normalization is performed as follows:

Enter the characteristic variables (the country's historical medal data: gold, silver, bronze; the number of athletes involved; the number of events involved), then substitute these datas into the formula.

$$Y = \{Y_1, Y_2, \cdots, Y_N\} \tag{1}$$

Where: $Y = \{G_i, S_i, B_i\}$ represents the number of gold, silver and bronze medals won by the i-th country in a particular year of the Olympic Games.

2.1.2 Feature selection

Select features that affect the total medal table, such as the number of historical medals, number of athletes, the number of events participated in by each country and the advantage programs etc[7]. For the correlation between the features is checked, avoiding multicollinearity.

2.1.3 Modeling establishment

In order to predict the number of gold medals and the total medal table of each country in 2028, this paper chooses to use a linear regression model. The linear regression model assumes a linear relationship between the number of medals and a set of characteristics. The regression model is set up as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \tag{2}$$

where: Y is the predicted number of gold medals or total medals; X_1,X_2,...,X_n are the characteristic variables, such as the number of historical medals, the number of athletes, the number of events, the infrastructure of each country etc; β_0 is the intercept term, which denotes the baseline number of medals when all the characteristic variables are zero; β_1, β_2,..,β_n is the regression coefficient, reflecting the degree of influence of each characteristic variable on the number of medals; ϵ is the error term, indicating random fluctuations and unexplained parts of the regression model. The magnitude of the regression coefficients reflects the degree of influence of each feature on the number of gold medals or the total number of medals. The regression coefficients were estimated from the training dataset. The aim is to minimize the error between the predicted and actual values.

2.1.4 Linear regression method

In this model, it is assumed that the number of medals (Y) is a linear relationship determined by the combination of a number of characteristics. For example historical number of medals and number of athletes etc. To estimate the regression coefficients, the ordinary least squares (OLS) method is used. This method estimates the regression coefficients by minimizing the error squared between the predicted and actual values.

Assume there are N training samples, and each sample contains n feature variables. The number of medals for each sample is denoted as y_i and the corresponding value of the feature variable is $X_{1i}, X_{2i}, \cdots, X_{ni}$.The goal is to minimize the objective function.

$$\text{miminize} \sum_{i=1}^{N} (y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}))^2 \tag{3}$$

Where: y_i is the actual number of medals; $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_n X_{ni}$ is the predicted value of the model. By minimizing the above objective function, it is possible to estimate the regression coefficients $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$.

To solve this optimization problem, the minimum value of this objective function is solved, usually by gradient descent or regular equations. The solution to the regular equation is :

$$\beta = (X^T X)^{-1} X^T Y \tag{4}$$

Where: $\beta = (\beta_0, \beta_1, \beta_2, \cdots, \beta_n)^T$ is the estimate of the regression coefficients; X is an N × (n + 1) matrix, where each row represents the eigenvector of a training sample and the first column is corresponding to the intercept term; Y is an N × 1 vector containing the actual medals of all the training samples; and X^T is the transpose matrix of X.

By solving this formal equation, estimates of the regression coefficients are obtained, and a prediction model is developed.

2.1.5 Calculation of expected number of gold medals and total medals

Input Characterization Variables. Bring to Model. Calculate for each country prediction of Gold Medal Count. And use a fitted model and known gold medal counts to predict the number of medals and their uncertainty intervals for each country at the 2028 Lod Angeles Olympics so that it can make sure the accuracy of the results. Ultimately, based on the predicted medal table, select the total medal count prediction interval as [-23,23] and the gold medal count prediction interval as [-9,9].

2.1.6 Construction of prediction intervals

Based on the nature of the linear regression model, combined with random effects, the predictive distribution of the number of gold medals in each country was generated, and the 95% prediction interval was extracted, and it was located in the interval $[\mu-3\sigma,\mu+3\sigma]$.

2.2 Model Evaluation

The goodness of a regression model is usually assessed by several indicators:

Coefficient of determination R^2: indicates the proportion of variability explained by the model, $R^2 \in [0,1]$, closer to 1 means better model fit.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{5}$$

Where: y_i is the predicted value and $\bar{y}_i$ is the average value of the sample.

Mean Square Error (MSE): indicates the squared mean of the error between the predicted and actual values, the smaller it is the better the model predicts.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{6}$$

Residual analysis: check whether the residuals (the difference between actual and predicted values) conform to a normal distribution and analyze whether there is a system errors of a sexual nature.

These evaluation metrics allow us to judge the predictive effectiveness of the regression model and further optimize the model to ensure that it is suitable for predicting the number of gold medals and the total number of medals for each country in 2028.

2.3 Analysis of Results

The regression analysis allows us to derive the degree of influence of each characteristic variable on the number of medals. The positive and negative regression coefficients indicate whether the relationship between the characteristics and the number of medals is positive or negative, and the absolute value of the coefficients indicates the magnitude of their influence on the number of medals. Through this prediction model, the paper obtained a regression coefficient of 0.7537, indicating that the number of historical medals has a strong impact on the number of gold medals and total medals.

The prediction result will give the number of total medals or total medals for each country in the 2028 Olympics, as in Figure 1,showing the predicted number of total medals for the 2028 Olympics, with the United States of America (USA) and China(CHN) showing a significant lead[8], which has a much higher number of predicted gold medals than any other country, followed by China (CHN) and Japan (JPN), demonstrating the strong competitiveness of these countries in the Olympics. Other countries such as Australia (AUS), France (FRA) and Great Britain (GBR) also demonstrated solid competitive performances. The calculation then yields a prediction interval of [-9,9] for the number of gold medals and [-23,23] for the total number of medals. And from the table, the total number of medals for the US is [90,136], and the prediction interval for China is [63,109], indicating a high degree of uncertainty in the results. So the prediction intervals may be strongly influenced by a variety of factors such as changes in team composition, athlete health, and training conditions. And based on historical data and predictions medal standings. It is possible to analyze those countries that progress or regress. And ten countries that are likely to progress or regress were selected, as shown in Figures 2 and 3. From a sociological perspective economic powerhouses and countries with long sporting traditions continue to dominate at the Olympics[9]. The United States will have 43 gold medals at the 2028 Olympics, an increase of 3 medals from 2024, and 113 total medals, a decrease of 13 medals from 2024; Because China has great competitiveness in traditional excellent events such as diving[10], it ranks the top in terms of gold medals and total medals. According to the figure, China will have 38 gold medals at the 2028 Olympics, a decrease of 2 medals and a total of 96 medals, an increase of 6 medals compared to 2024. These countries are traditional sports powerhouses with

little fluctuation in the number of gold medals and total medals. Their gold and total medal counts are likely to improve. On the contrary Fiji, Peru's gold medal count and total medal count may decrease, Fiji may decrease by 25 medals, but Fiji won only 1 medal in 2024 and may not win a medal in 2028, and these countries are generally economically backward and weak sports countries.

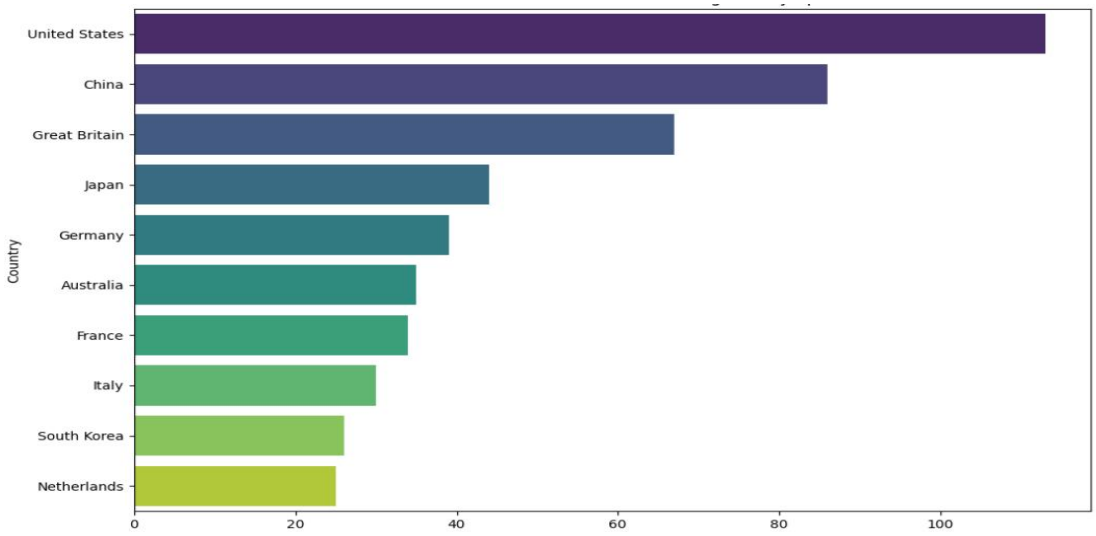

Figure 1 The Top Ten Countries in terms of Total Medals in the Projected 2028 Medal Table

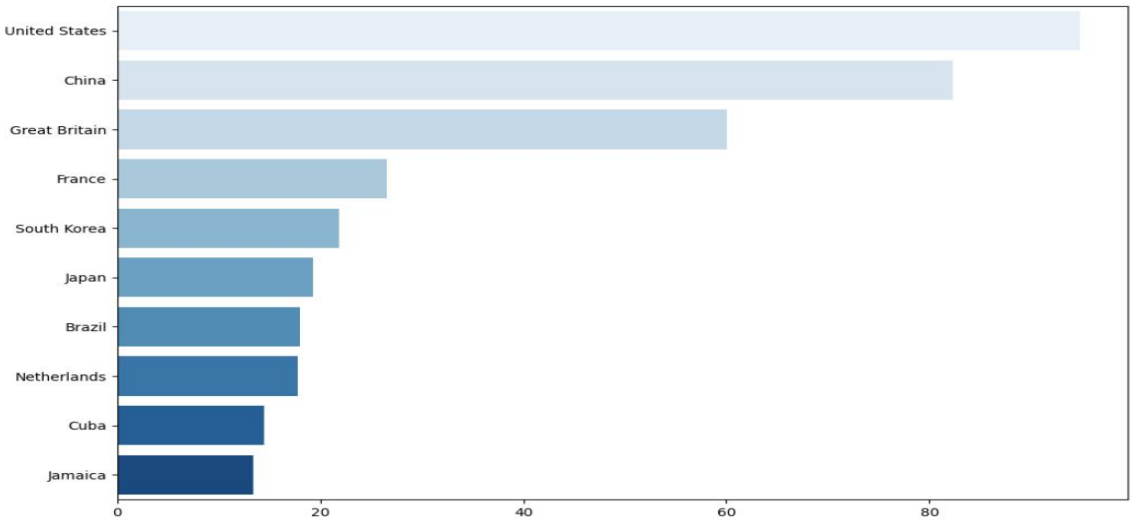

Figure 2 The Top Ten Countries Selected for Possible Progress in the Olympic Medal Table

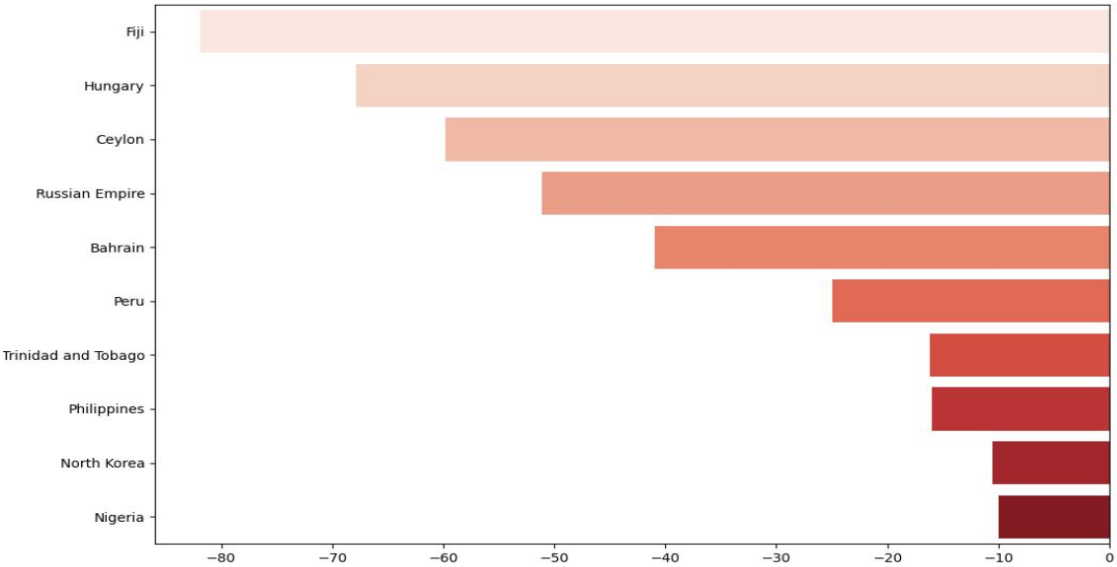

Figure 3 The Ten Countries Selected for Possible Regression in the Olympic Medal Table

3 THE FIRST-TIME AWARD-WINNING COUNTRIES BASED ON LOGISTIC MODEL

3.1 Research Methodology

According to the historical medal list to select those countries that have never won the Olympic Games as the independent variable, select the probability of winning the award as the dependent variable, to establish a logistic regression-based binary classification model to explore the likelihood of winning the award, the interval is [0,1], if the probability of winning the award of a country is closer to 1, the greater the likelihood of winning the award, the reverse is not the case.

3.2 Modeling Selecting

In order to make categorical predictions, logistic regression model is chosen in this paper. Logistic regression is a commonly used binary classification model that makes predictions by calculating the probability of an event occurring. Specifically for this study, the objective of this paper is to predict whether a country will be able to win a medal or not, and in particular whether the country will be able to win a medal for the first time in the 2028 Olympic Games. The mathematical expression of the logistic regression model is as follows.

$$\mathrm{P(Medal)} = \frac{1}{1+\mathrm{e}^{-(\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\cdots \beta_n \mathrm{X}_n)}} \tag{7}$$

Where: P(medal) represents the probability of a country winning a medal, i.e. the likelihood of that country winning a medal in the 2028 Olympics. Since it is a probability value, the value of P(medal) must lie between 0 and 1; $X_1,X_2,\ldots,X_n$ are the characteristic variables affecting whether or not the country can win a medal. These characteristics can be the number of athletes in the country, the number of participating sports, historical performance, economic level etc, which together determine the probability of the country to win a medal; $\beta_0,\beta_1,\ldots,\beta_n$ are the regression coefficients, which indicate the degree of influence of the characteristics on the winning of medals. Through the training of the model, we can estimate the values of these coefficients; e is the base of the natural logarithm, which is used to ensure that the output probability value of the model is always between 0 and 1, in line with the definition of probability.

The goal of this logistic regression model is to learn the regression coefficients $\beta_0, \beta_1 ,\ldots,\beta_n$ from the given training data. The process of learning the regression coefficients is achieved by maximizing the likelihood function. The purpose of maximizing the likelihood function is to make the probability values predicted by the model as close as possible to the actual observed labels.

3.2.1 Maximum likelihood estimation

In logistic regression, the regression coefficients are estimated by maximizing the likelihood function. The likelihood function represents the probability of observing the current data given the characteristic data. Assuming that there are m samples with label 3 and feature X_i ,the likelihood function can be expressed as follows.

$$L(\beta_0,\beta_1,\ldots\ldots,\beta_n) = \prod_{i=1}^{m} \mathrm{P}(\mathrm{y_i} \mid \mathrm{X_i}) \tag{8}$$

where $\mathrm{P}(y_i|X_i)$ denotes the probability that the label yiof sample i is 1. Since this paper is a binary classification problem, label i takes the value of 0 or 1, which indicates whether the country wins a medal or not, respectively. Therefore, $\mathrm{P}(\mathrm{y_i}|\mathrm{X_i})$ can be written as.

$$\mathrm{P}(\mathrm{y_i} \mid \mathrm{X_i}) = \mathrm{P(Medal)}^{\mathrm{y_i}} (1-\mathrm{P(Medal)})^{(1-\mathrm{y_i})} \tag{9}$$

The probability of a sample i-th is P(medal), if its label $y_i = 1$(namely the country won a medal); if $y_i = 0$(namely the country did not win a medal), the probability is $1 - \mathrm{P(Medal)}$.

In order to simplify the calculation and improve the numerical stability, we usually take the logarithm of the likelihood function to get the log-likelihood function. The log-likelihood function is expressed as.

$$\ell(\beta_0,\beta_1,\cdots\cdots,\beta_n) = \sum_{i=1}^{m} [\mathrm{y_i log(P(Medal)) + (1-y_i) log(1-P(Medal))}] \tag{10}$$

By maximizing the log-likelihood function, we are able to obtain the regression coefficients $\beta_0 ,\beta_1,\ldots,\beta n$. The goal of maximizing the log-likelihood function is to make the predicted probabilities of the model as consistent as possible with the actual labels. This process usually uses optimization algorithms (like gradient descent) to find the optimal regression coefficients.

3.2.2 Solving for regression coefficients

The regression coefficients are solved by maximizing the log-likelihood function. To solve these coefficients, numerical optimization methods are usually used. In logistic regression, commonly used optimization methods include gradient descent and Newton's method. The gradient descent method gradually finds the parameters that maximize the log-likelihood function by calculating the gradient of the log-likelihood function and updating the regression coefficients at each iteration step.

Specifically, the gradient descent method works by calculating the partial derivatives of each regression coefficient and updating the regression coefficients based on the derivative values. At each iteration, the regression coefficients are

adjusted in a direction that causes the log-likelihood function to increase. The iterative process continues until the log-likelihood function converges, until the regression coefficients no longer change significantly.

3.2.3 Model solving

In order to implement the logistic regression model and predict the probability that a country that has not yet won a medal will win a medal for the first time, the input eigenvectors are passed through the model and the paper uses a binary classification model, by fitting the probability of non award-winning country to a distance of 0 or 1,then can conduct whether they can or can't win the first medal in 2028 Olympic.

3.3 Analysis of Results

After obtaining the model results, the probability of each yet-to-be-awarded country winning a medal can be output. With the model, we calculate the probability that each country that has not yet won a medal will win its first medal in 2028. There are 115 countries that have not yet won a medal, and the model predicts that 10 of them will win their first medal in 2028, and the probability for each of them is more than 0.5. This means that there is a high level of confidence that these 10 countries will break through history and reach the podium for the first time. Through this prediction method, we can provide data support to the National Olympic Committees to help them develop more targeted Olympic strategies. In addition, through further analysis of the model, we can identify the important factors that affect a country's ability to win medals, such as the quality of athletes, the number of events entered, and historical performance. These factors will help countries to make more precise adjustments in future Olympic Games. For this purpose, we have selected ten countries that are most likely to win a medal for the first time. As shown in Table1, the Republic of Armenia has the highest probability of winning a prize for the first time at the 2028 Olympic Games.

Table 1 The Top Countries most likely to Win a Medal for the First Time

COUNTRY	POSSIBILITY
Armenia	0.7956
Bahamas	0.7753
Azberbaijan	0.7485
Algeria	0.7165
Bahrain	0.6647
Albania	0.6381
Tonga	0.6257
Barbados	0.6052
Peru	0.5860
Afghanistan	0.5765

4 ESTIMATION OF R2 AND MSE FOR PREDICTION OF MULTIPLE LINEAR REGRESSION MODELS

In conducting the performance evaluation of the linear regression model for the prediction of the number of medals for each country at the 2028 Summer Olympics in Los Angeles, an impressive set of statistical metrics were obtained, which emphasize the high accuracy and reliability of the model. Specifically, the mean square error (MSE) of the model was 0.85, showing a small prediction error, implying a low mean squared difference between predicted and actual values. This low level of error indicates that the model performs well in predicting the number of medals for each country with proper error control. The Mean Absolute Error (MAE) of 0.91 further confirms the model's ability to maintain consistency across data points, reflecting a low mean absolute deviation between predicted and actual values. This is particularly important because it is directly related to the usefulness and reliability of the prediction results, and the low MAE value indicates that the model has less error in practical applications and more accurate prediction results. The coefficient of determination (R^2) of the model reaches 0.99, which is almost perfect performance, and almost all the data variations can be explained by the model. This high R^2 value not only demonstrates the statistical superiority of the model, but more importantly, it shows that the model is able to capture and explain the various factors affecting the number of medals extremely effectively, ensuring a high degree of accuracy and explanatory power in the prediction results.

5 CONCLUSION

The Olympic Games, as the largest sports event, has always been a focus of attention in predicting the number of medals won. The paper uses a multiple linear regression model to predict the medal table and analyze the progress and regression of the countries, and it estimate MSE and R^2, and R^2 is 0.99 while MSE is 0.91 by calculating. In addition, logistics regression model and binary classification model are used to explore the winning situation of countries that have not won the first medal. The multiple linear regression model can quantitatively analyze multiple factors that affect the number of medals, clarifying the degree and direction of each factor's impact on the number of the medals won. And multiple linear regression model can be combined with other prediction models(such as random forests) to comprehensively utilize the advantages of different models and improve the accuracy and reliability of predictions. In the future, based on this model, multiple analysis can be conducted to predict which Olympic events have a high probability of China winning the championship. This can provide valuable information for relevant sports organizations

to better understand the possible direction of future Olympic games, so that the country can formulate more scientific and effective training and preparation strategies.

COMPETING INTERESTS

The author has no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Abel Sowal. Olympic medal prediction is not just for "presence". Economic and Social Management Science, 2016, (20): 77.
[2] Nagpal Prince, Gupta Kartikey, Verma Yashaswa, et al. Paris Olympic(2024) Medal Tally Prediction. Banaras Hindu University,Varanasi, India, Lecture Notes in Networks and Systems, 2023, (662): 249-267.
[3] Cha Zheng. Economic Reflections on Olympic Medals. Shandong Radio and Television University, Social Science II Series, 2020, (05): 97.
[4] Mallappa Uday, Gangwar Pranav, Khaleghi Behnam, et al. TermiNETor:Early Convolution Termination for Efficient Deep Netural Networks. 2022 IEEE 40th International Conference on Computer Design (ICCD), Olympic Valley, CA, USA, 2022: 635-643. DOI: 10.1109/ICCD56317.2022.00098.
[5] Deng Rongrong, Fan Qingmin, Zhang Yinkai, et al. Analysis of Technical Evaluation of Male Boxing Athletes Based on Entropy and Multiple Linear Regression Model-Taking Excellent 81kg Athletes at the Tokyo Olympics as an Example. School of Competitive Sports,Beijing Sport University, Social Science II Series, 2023, (11): 4744-4746.
[6] Miyoshi Takemasa, Amemiya Arata, Otsuka Shigenori, et al. Big Data Assimilation:Real-time 30-second-refresh Heavy Rain Forecast Using Fugaku during Tokyo Olympics and Paralympics. Meteorological Research Institute,Tsukuba, Japan, Association for Computing Machinery, Inc, 2023, (12): 8.
[7] Wang Yongjun, Chen Hong, Yang Yongfen. Fuzzy relationship,mediating variables,and dynamic models:A study on the trickle down effect of sports participation in heritage, School of Physical Education. Chongqing Technology and Business University,School of Management,Tianjin Sport University, Kunming University, Social Science II Series, 2021, (7): 28-30.
[8] Gupta Krishon Gopal, Arora Aditi. Olympic Data Analysis: Uncovering New Insights into Athletic Performance and Competition. ABES Engineering College, Ghaziabad, India, Grenze Scientific Society, 2024, (2): 4312-4319.
[9] Li Luyan, Gao Yong. Understanding of Beijing Winter Olympics Volunteers from a Sociological Perspective. School of Physical Education,Henan University of Science and technology, Social Science II Series, 2022, (10): 216-217.
[10] Ning Yixia. Observing the characteristics of China’s Competitive Sports Strength from the Tokyo Olympics. Shenzhen University, Social Science II Series, 2022, (09): 36-38.